VIDEO-BASED, OCCLUSION-ROBUST MULTI-VIEW STEREO USING INNER-BOUNDARY DEPTHS OF TEXTURELESS AREAS

Jian Wei, Shigang Wang, Yan Zhao

College of Communication Engineering, Jilin University, Changchun 130012, China

ABSTRACT

Occlusions and poor textures are two main problems in multi-view stereo reconstruction. This paper presents a video-based solution to address both challenges in depth estimation. We focus on reconstructing accurate inner boundaries of visible textureless areas, particularly for occluded background, by leveraging the reliable depths of object edges. This is done by efficiently respecting two local cues with complementary advantages, *i.e.* smoothness and density of recovered surfaces. The inner-boundary depths are finally utilized to infer dense geometry without wrong connections between objects. This method only relies on low-level techniques, *e.g.* intra-view interpolation and inter-view propagation of depths. Experiments indicate its superiority in terms of both depth discontinuities near object silhouettes and surface smoothness in homogeneous regions compared to the state of the art.

Index Terms— Multi-view stereo, monocular video, occlusion, poor textures, 3D reconstruction

1. INTRODUCTION

The goal of Multi-View Stereo (MVS) is to reconstruct dense three-dimensional (3D) scene geometry from a set of images with known camera calibration and poses. It plays a major role in numerous applications including object segmentation, robotics, novel view synthesis, stereoscopic displays, and virtual/augmented reality. Conventional MVS methods [1, 2, 3] focus on dealing with images that are captured at sparsely distributed viewpoints. Most of them calculate depth by matching two-dimensional (2D) patches among several views, and thus usually get blurred depth discontinuities at object edges. Moreover, severer occlusions and poorly textured surfaces in the scene normally lead to unreliable cross-view correlation, resulting in wrong depth estimates.

This paper aims at solving the above problems by using a casually shot high-frame-rate video. This MVS strategy has attracted much attention in recent years for two primary reasons: First, video acquisition is simpler, faster, and capable of ensuring denser coverage of a scene than taking photographs. Second, the densely sampled input provides sufficient information for accurate 3D scene reconstruction.

Various video-based MVS methods have been proposed. To overcome weak textures, Garg *et al*. [4] introduced a non-convex energy function with a total variation regularization. It was enhanced in [5] via a convex coherent depth field energy for fewer operations and easier parallelism. Kumar *et al*. [6] exploited superpixels for the scenes containing mainly planar surfaces. Occlusions were handled in [7] by a template calculated from at least one view. In [8], a shape prior obtained from a few unoccluded views was used as a constraint to recover occluded regions. Quality can be improved by leveraging the hierarchical features from deep convolutional neural networks [9] or a more advanced deep ordinal regression network [10]. Xu *et al*. [11] integrated conditional random fields into the deep architectures. These techniques have also been adopted to address the same problems in depth estimation of multi-view photographs [12, 13] and light fields [14, 15].

Our work is inspired by two previous researches. In [16], the diffused depths of well-textured edges are used as the initial guess of interior homogeneous surfaces. Then, erroneous estimates are corrected for each view by detecting the connections that occlude the edges of other views, spreading depth errors, and improving surface coverage as well as multi-view consistency through cross-view depth propagation. The performance of this scheme is determined by the accuracy and completeness of edge depths. In [17], the reconstructed edges are separated into the ones belonging to different objects, and depth interpolation is done by using each individual segment to avoid occlusion errors. Since the obtained inner boundaries of background surfaces are interrupted near the foreground, wrong and inconsistent interpolants are still unavoidable.

Our paper borrows the idea of image-domain region growing from [16, 17], as it is easier to implement than the methods designed in 3D space [18, 19]. However, unlike [17], the core of our scheme is to reconstruct complete inner boundaries of visible weakly textured areas in individual views (Fig. 1(g)), some of which are view-dependent due to occlusions. Taking their depths as seeds, we are able to preserve the sharp surface discontinuities between fore- and background objects when interpolating depths, hence the occlusion reasoning of [16], which is sensitive to noises of edge depths, is not needed. We

This work was supported by the National Key Research and Development Plan of 13th Five-Year (2017YFB0404800), National Natural Science Foundation of China (NSFC) (61631009), and Fundamental Research Funds for the Central Universities (2017TD-19).

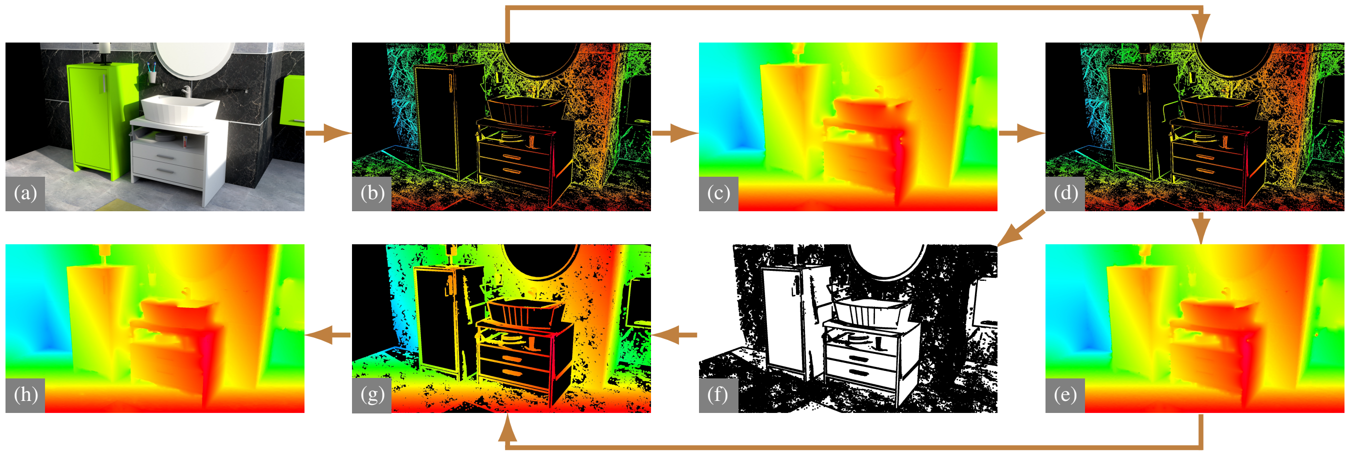

Fig. 1. Flowchart of the proposed approach. (a) One frame. (b) Edge depths of (a). (c) Interpolated depths from (b). (d) Refined edge depths. (e) Interpolated depths from (d). (f) Binary mask from expanded (d). (g) Inner-boundary depths of the textureless areas in (a). (h) Interpolated depths from (g).

do this by jointly considering the smoothness and density of 3D surfaces, similarly to the point cloud merging algorithm of [20] but without calculating the expensive implicit function. These two local cues have their respective advantages in reconstructing homogeneous surfaces and preserving object silhouettes. Our approach also contains an edge depth refinement stage for better recovery of dense scene geometry that is lacked in [16]. It requires no high-level technique, *e.g.* image over-segmentation [6], energy minimization [4, 5], and deep learning [9, 10, 11]. Instead, most operations are evaluated per pixel, thus supporting parallel execution on GPUs.

2. OUR PROPOSED METHOD

Fig. 1 illustrates the main procedures of our video-based MVS method. Given a monocular video with unconstrained camera trajectory, we first execute the Structure-from-Motion (SfM) approach of Resch *et al*. [21] to compute the camera extrinsics of all frames, select a dense subset of frames $\{\mathbf{I}_i\}$ for sufficient triangulation angles, and calculate the depths of edge pixels (the magnitude of color gradient computed with a 3×3 Sobel operator is larger than 2.5) for each $\mathbf{I}_i$ (Fig. 1(a)) at the single pixel level. We utilize the techniques proposed in [16] for frame sampling and edge depth estimation. Then, the obtained edge depth maps $\{\mathbf{D}_i^{\mathrm{e}}\}$ (Fig. 1(b)) are further refined to improve edge continuity (Figs. 1(c)$\sim$1(d)), and afterwards leveraged to derive the depth maps $\{\mathbf{D}_i^{\mathrm{b}}\}$ for the inner boundaries of interior textureless areas (Figs. 1(e)$\sim$1(g)). The dense depth maps $\{\mathbf{D}_i^{\mathrm{f}}\}$ (Fig. 1(h)) are finally obtained through per-view interpolation of $\{\mathbf{D}_i^{\mathrm{b}}\}$. Occlusions and lack of textures are solved by effectively combining the depth values that produce the smoothest and densest surfaces, as described below.

2.1. Edge depth refinement

Due to viewpoint change, some pixels corresponding to the object silhouettes have different magnitudes of color gradient in different frames, *e.g.* the cabinet edges under strong light in Fig. 1(a). Depth estimation possibly fails on these pixels (Fig. 1(b)), that may lead to wrong surface interpolants in our method (Fig. 6(d)). Therefore, it is necessary to fill in the edge gaps in advance. Because the missing 3D edges can be more easily reconstructed from the projections with prominent textures in the associated views, we enhance depth continuity for each $\mathbf{D}_i^{\mathrm{e}}$ by propagating depths across all views.

Considering that the depth sparsity probably causes depth propagation from distant objects to close-up objects, our refinement process consists of two steps. First, we implement the depth diffusion algorithm of [16] on each $\mathbf{D}_i^{\mathrm{e}}$, that fits the smoothest possible surface for homogeneous areas, obtaining $\mathbf{D}_i^{\mathrm{s}}$ (Fig. 1(c)). After that, we update $\mathbf{D}_i^{\mathrm{s}}$ by warping all $\{\mathbf{D}_j^{\mathrm{s}}\}$ to its view and picking the least distant depth at each pixel $\mathbf{p}$:

$$(\tilde{j},\tilde{\mathbf{q}}) = \arg\min_{(j,\mathbf{q})} \underset{j\to i}{\mathbf{T}}\left(\mathbf{D}_j^{\mathrm{s}}(\mathbf{q})\right), \quad \text{s.t. } \underset{j\to i}{\mathbf{P}}\left(\mathbf{D}_j^{\mathrm{s}}(\mathbf{q})\right) = \mathbf{p} \quad (1)$$

where $\underset{j\to i}{\mathbf{P}}\left(\mathbf{D}_j^{\mathrm{s}}(\mathbf{q})\right)$ and $\underset{j\to i}{\mathbf{T}}\left(\mathbf{D}_j^{\mathrm{s}}(\mathbf{q})\right)$ represent the pixel and depth obtained by projecting the pixel $\mathbf{q}$ from $\mathbf{D}_j^{\mathrm{s}}$ to the ith view, respectively. The refined $\mathbf{D}_i^{\mathrm{e}}$ is composed of the depths $\{\mathbf{D}_{\tilde{j}}^{\mathrm{s}}(\tilde{\mathbf{q}})\}$, where $\tilde{\mathbf{q}}$ is an edge pixel. The outliers are removed from $\mathbf{D}_i^{\mathrm{e}}$ by a 7×7 median filter, that only considers the neighboring pixels with small color difference (< 5) to preserve the localization of silhouettes. Fig. 1(d) demonstrates that the above pre-processing improves the edge depth completeness.

2.2. Textureless surface reconstruction

After refining $\mathbf{D}_i^{\mathrm{e}}$, we can yield a more accurate $\mathbf{D}_i^{\mathrm{s}}$ via depth interpolation (Fig. 1(e)). However, as shown in Fig. 2(c), only considering surface smoothness induces some incorrect interpolants in front of the real scene geometry. The reason lies in that the edges (on the foreground side) of some background homogeneous regions are occluded in this view. We handle these errors by diffusing the inner-boundary depths for each

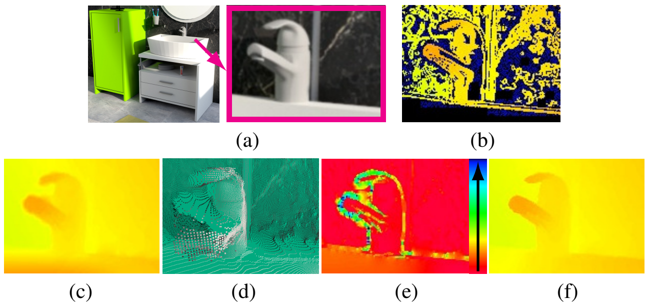

Fig. 2. Benefit of respecting surface density. (a) Image area. (b) Edge depths of (a). (c) Interpolated depths from (b). (d) 3D points produced from (c) (rotated for better visualization). (e) Surface densities of (c), decreasing in the direction of the arrow. (f) Angularly propagated depths using the density cue.

visible poorly textured area. The inner boundaries are reconstructed by additionally respecting the cue of surface density.

2.2.1. *Benefit of respecting surface density*

Fig. 2(d) visualizes the 3D points reconstructed using Fig. 2(c). It can be seen that, most of the surfaces connecting fore- and background edges are heavily slanted for the corresponding view, thus having much lower density than others. In view of this, one way of correcting these erroneous depths in $\mathbf{D}^{\mathrm{s}}_i$ is to propagate the estimates, that produce the highest-density surfaces, across all $\{\mathbf{D}^{\mathrm{s}}_j\}$. To this end, we measure the density of per-view recovered surfaces by using the maximum distance between the 3D points calculated from each pixel $\mathbf{p}$ and its 4-neighbors (Fig. 2(e)). With the surface densities of $\mathbf{D}^{\mathrm{s}}_j$, $\mathbf{S}_j$, the aforesaid angular depth propagation is formulated as:

$$(\tilde{j}, \tilde{\mathbf{q}}) = \arg\min_{(j,\mathbf{q})} \mathbf{S}_j(\mathbf{q}), \quad \text{s.t. } \underset{j\to i}{\mathbf{P}}\left(\mathbf{D}^{\mathrm{s}}_j(\mathbf{q})\right) = \mathbf{p} \tag{2}$$

The estimates $\{\mathbf{D}^{\mathrm{s}}_{\tilde{j}}(\tilde{\mathbf{q}})\}$ comprise the resulted depth map. By comparing Figs. 2(c) and 2(f), we can obviously observe that the depth discontinuities near object silhouettes are better preserved by taking into account surface density.

2.2.2. *Inner-boundary-based depth calculation combing surface smoothness and density*

Fig. 3 gives another example. As depicted in Fig. 3(c), simply selecting the densest surfaces among all views probably lead to layered geometry in weakly textured regions, where the 3D structure is expected to be flat (Fig. 3(b)). We thereby propose to avoid these artifacts by merely depending on the cue of surface density at the inner boundaries of homogeneous areas, where the depths obtained in such a way are usually reliable. Afterwards we infer dense geometry for interior areas under the smooth-surface assumption.

Specifically, Fig. 4(b) shows the edge depths of an image region where occlusions occur. We note that, near the surface demarcations, only the boundaries of foreground objects can

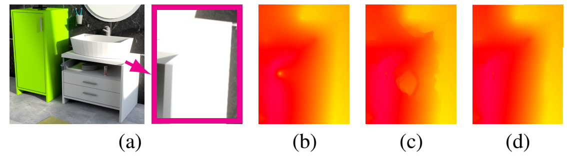

Fig. 3. Shortcoming of respecting surface density. (a) Image area. (b) Interpolated estimates from edge depths. (c) Angularly propagated depths using the density cue. (d) Interpolated estimates from inner-boundary depths of textureless areas.

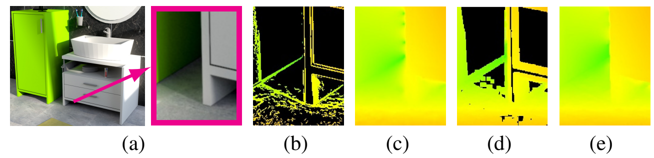

Fig. 4. Depth estimation combining the smoothness and density cues of surfaces. (a) Image area. (b) Edge depths of (a). (c) Interpolated depths from (b). (d) Inner-boundary depths of textureless areas. (e) Interpolated depths from (d).

Fig. 5. Sampled images of (a) Boxes and (b) Building.

be reconstructed in the edge depth estimation stage. For this reason, our algorithm first extracts the inner boundaries of all homogeneous areas using morphology operation. In particular, we expand the binary mask (set to 1 for a pixel if it has a depth value, and 0 otherwise) of each $\mathbf{D}^{\mathrm{e}}_i$ by ω pixels, obtaining the mask $\mathbf{M}_i$ (Fig. 1(f)). The density-first inter-view depth propagation step of Eq. 2 is subsequently performed, but limited solely to the pixels $\{\mathbf{p}\}$ satisfying $\mathbf{M}_i(\mathbf{p}) = 1$. With the derived sparse inner-boundary depth map $\mathbf{D}^{\mathrm{b}}_i$ (Fig. 1(g)), we use the smoothness-priority intra-view depth interpolation of [16] to get the final dense depth map $\mathbf{D}^{\mathrm{f}}_i$ (Fig. 1(h)).

The above-mentioned width of edge expansion, ω, plays the role of adjusting the balance between the cues of surface smoothness and density. If $\omega = 0$, $\{\mathbf{D}^{\mathrm{s}}_i\}$ would be the final output of the presented method. A larger value assigns more weight to the depths generating compact surfaces. $\omega = 5$ is used in our implementation.

As illustrated in Fig. 4(d), our approach is able to attain accurate inner-boundary depth estimation, especially for the partially occluded background regions near the foreground. Integrating both cues recovers sharper object silhouettes than only considering surface smoothness (compare Figs. 4(c) and 4(e)) and more flat geometry than only considering surface density (compare Figs. 3(c) and 3(d)).

3. EXPERIMENTS

Our experiments used three videos with large textureless areas and arbitrary camera trajectories: one synthetic scene *Bathroom* created in Blender software (Fig. 1), as well as two real-

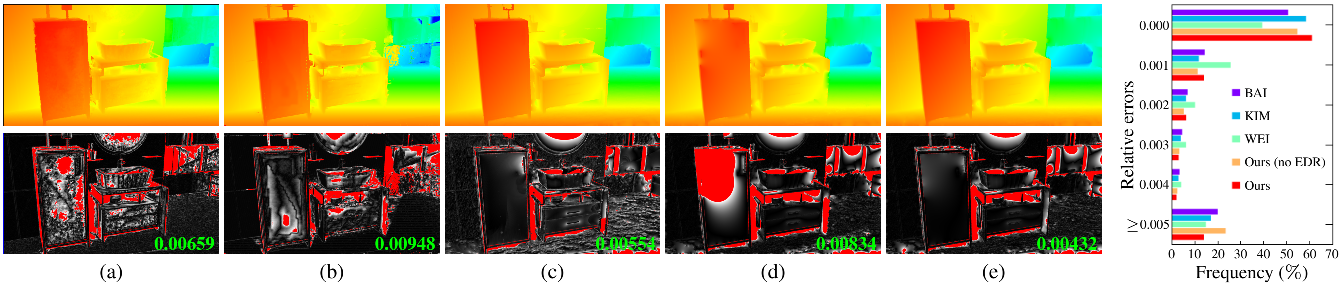

Fig. 6. Depth maps (top), relative error (RE) maps against the ground truth (bottom), and RE distributions (right) obtained by (a) BAI, (b) KIM, (c) WEI, and ours (d) without and (e) with the edge depth refinement step (EDR, Section 2.1). In the RE maps, the blue pixels have no depth, the red have a ER larger than 0.01, the REs between 0 and 0.01 are marked gray 0 to 255, and the mean RE of each method is labeled in the lower right corner.

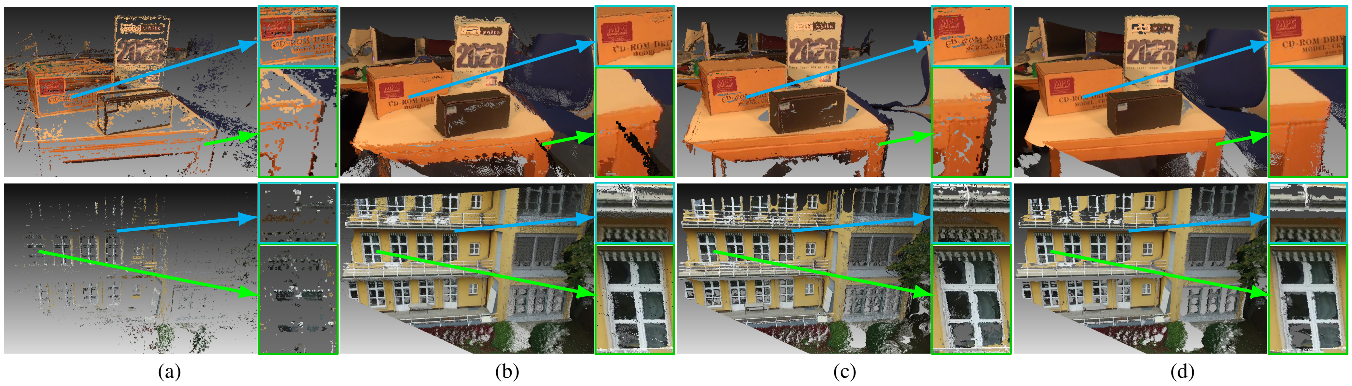

Fig. 7. Point clouds of real-world scenes reconstructed from the images in Fig. 5 by (a) BAI, (b) KIM, (c) WEI, and (d) ours.

world scenes *Boxes* and *Building* without known geometry (Fig. 5). 100 frames were pre-selected from each dataset, and all images have resolution of 1920×1080.

Three depth map calculation algorithms were compared: the PatchMatch-based approach *BAI* [1] for photographs, the light field reconstruction work *KIM* [22] via depth sweeping (1024 hypotheses), and the video-oriented scheme *WEI* [16]. Evaluations were run on multithreaded CPUs, and NVIDIA GeForce GTX 680 (KIM) and 1080 Ti (others) GPUs.

Fig. 6 quantitatively compares the depth maps computed using different methods. According to the relative error (RE) maps, BAI and KIM both generate uneven surfaces in most homogeneous areas. Although WEI does not suffer from such artifacts, like ours, the RE distributions show that it produces the least high-accuracy depth values (RE < 0.001) while ours yields the most. Besides, the proposed approach generates a minimum amount of estimates with RE ≥ 0.005. It also appears that the refined edge depths improve our reconstruction quality. As indicated by the mean REs, our estimates have the highest overall precision.

Fig. 7 assesses the ability of various methods on recovering realistic geometry. Since noises and compression artifacts in the captured images caused many wrong estimates, we removed the isolated points with large distance from neighborhoods for all schemes. Our obtained point clouds are significantly denser than BAI's, cleaner than KIM's, and smoother than WEI's, even for the areas lacking textures. Furthermore, the surface discontinuities of our results align best with the object silhouettes in contrast to all other algorithms.

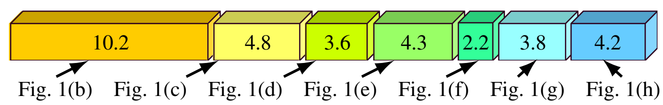

Fig. 8. Average runtime (sec.) of the proposed pipeline for calculating individual immediate results in Fig. 1.

Fig. 8 provides the average runtime for reconstructing one view in each step of our pipeline. The edge depth estimation takes the most time but is beyond the scope of this paper. Even so, the total runtime of 33.1 secs. is still less than those of BAI (71.8 secs.), KIM (122.2 secs.), and WEI (34.1 secs.).

4. CONCLUSIONS

This paper presented an efficient, video-based MVS method that can handle occlusions and weak textures. Our main contribution is the utilization of edge depths to reconstruct the inner boundaries of visible homogeneous surfaces. This goal is attained by making the cues of surface smoothness and density complement each other. The results demonstrate that the proposed approach outperforms the state of the art in preserving both object silhouettes and surface smoothness.

5. REFERENCES

[1] C. Bailer, M. Finckh, and H.P.A. Lensch, "Scale robust multi view stereo," in *ECCV*, 2012.

[2] J. Wei, B. Resch, and H. P. A. Lensch, "Multi-view depth map estimation with cross-view consistency," in *BMVC*, 2014.

[3] P. H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J. B. Huang, "Deepmvs: learning multi-view stereopsis," in *IEEE CVPR*, 2018.

[4] R. Garg, A. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *IEEE CVPR*, 2013.

[5] V. Golyanik, T. Fetzer, and D. Stricker, "Introduction to coherent depth fields for dense monocular surface recovery," in *BMVC*, 2017.

[6] S. Kumar, Y. Dai, and H. Li, "Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames," in *IEEE ICCV*, 2017.

[7] R. Yu, C. Russell, N. D. F. Campbell, and L. Agapito, "Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video," in *IEEE ICCV*, 2015.

[8] V. Golyanik, T. Fetzer, and D. Stricker, "Accurate 3d reconstruction of dynamic scenes from monocular image sequences with severe occlusions," in *IEEE WACV*, 2017.

[9] Y. Kuznietsov, J. Stckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *IEEE CVPR*, 2017.

[10] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *IEEE CVPR*, 2018.

[11] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *IEEE CVPR*, 2018.

[12] X. Wang, M. Salzmann, F. Wang, and J. Zhao, "Template-free 3d reconstruction of poorly-textured nonrigid surfaces," in *ECCV*, 2016.

[13] I. Cherabier, J. L. Schönberger, M. R. Oswald, M. Pollefeys, and A. Geiger, "Learning priors for semantic 3d reconstruction," in *ECCV*, 2018.

[14] X. Jiang, M. L. Pendu, and C. Guillemot, "Depth estimation with occlusion handling from a sparse set of light field views," in *IEEE ICIP*, 2018.

[15] H. Schilling, M. Diebold, C. Rother, and B. Jähne, "Trust your model: Light field depth estimation with inline occlusion handling," in *IEEE CVPR*, 2018.

[16] J. Wei, B. Resch, and H. P. A. Lensch, "Dense and occlusion-robust multi-view stereo for unstructured videos," in *Conf. Comput. and Robot Vis. (CRV)*, 2016.

[17] Z. Ma, Z. Cen, and X. Li, "Depth estimation algorithm for light field data by epipolar image analysis and region interpolation," *Appl. Opt.*, vol. 56, no. 23, pp. 6603–6610, 2017.

[18] A. Usumezbas, R. Fabbri, and B. B. Kimia, "The surfacing of multiview 3d drawings via lofting and occlusion reasoning," in *IEEE CVPR*, 2017.

[19] T. Holzmann, M. Maurer, F. Fraundorfer, and H. Bischof, "Semantically aware urban 3d reconstruction with plane-based regularization," in *ECCV*, 2018.

[20] S. Fuhrmann and M. Goesele, "Floating scale surface reconstruction," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–11, 2014.

[21] B. Resch, H. P. A. Lensch, O. Wang, M. Pollefeys, and A. Solkine-Hornung, "Scalable structure from motion for densely sampled videos," in *IEEE CVPR*, 2015.

[22] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, pp. 73:1–12, 2013.